MOVING

Deliverable 6.2: Data Management Plan

Chrysa Collyda, Vasileios Mezaris, Sabrina Herbst, Paul Grunewaldm, Thomas Köhler, Angela Fessl, Ahmed Saleh, Till Blume, Falk Böschen, Ansgar Scherp, Markel Vigo, Tobias Backes, Peter Mutschke, Andrzej Skulimowski

30/09/2016

Work Package 6: Project management

TrainiG towards a society of data-saVvy inforMation prOfessionals to enable open leadership INnovation

Horizon 2020 - INSO-4-2015

Research and Innovation Programme

Grant Agreement Number 693092

Dissemination level	*PU*
Contractual date of delivery	*30/09/2016*
Actual date of delivery	*30/09/2016*
Deliverable number	*D6.2*
Deliverable name	*Data management plan*
File	*MOVING_D6.2_v1.0*
Nature	*Report*
Status & version	*Final v1.0*
Number of pages	*39*
WP contributing to the deliverable	*WP6*
Task responsible	*CERTH*
Other contributors	*All*
Author(s)	*Chrysa Collyda, Vasileios Mezaris* *CERTH* *Sabrina Herbst, Paul Grunewaldm, Thomas Köhler* *TUD* *Angela Fessl* *KC* *Ahmed Saleh, Till Blume, Falk Böschen, Ansgar Scherp* *ZBW* *Markel Vigo* *UMAN* *Tobias Backes, Peter Mutschke* *GESIS* *Andrzej Skulimowski* *PBF*
Quality Assessors	*Sebastian Gottfried, Paul Grunewald, TUD*
EC Project Officer	*Hinano SPREAFICO*
Keywords	*Data Management Plan*

Table of contents

Executive Summary

This deliverable presents the Data Management Plan of the MOVING project. In particular, it describes in detail the adopted management policy for the datasets that will be collected, processed or generated by the project. The utilized approach: (a) identifies which data and how they will be exploited or made publicly accessible so as to maximize their reuse potential, (b) specifies how these data will be curated and preserved, to support their reuse, and (c) identifies any data that should not be made publicly available and measures to be taken for their safe-keeping.

The European Commission (EC) has defined a number of guidelines / requirements for maximizing scientific data's reuse potential, via making them easily discoverable, intelligible, usable beyond the original purpose for which they were collected and interoperable to specific quality standards. Using these guidelines as a basis, we apply the methodology that is outlined in Section 1. According to this approach, for each dataset we specify: (a) its name (based on a standardized referencing approach), (b) its description, (c) the utilized standards and metadata, (d) the applicable data sharing policy and (e) the intended actions for its archiving and preservation. Further explanation regarding the information that needs to be considered and reported for each one of the above points is given in Sections 1.1. to 1.5. Subsequently, based on this methodology, Section 2 lists and describes the datasets of the MOVING project in a per-work-package-basis (Sections 2.1 to 2.5). It should be noted that WP4 is focused on the software development of the platform, thus is not expected to generate any dataset, or use datasets other than those already specified by the other work-packages. The concluding Section 3 briefly summarizes the information reported in the deliverable.

In the present document we discuss 29 different datasets, some of them being preexisting ones that will be used in MOVING, while others being new datasets that will be created during the lifetime of the project. Most of the dataset are or will become publicly available, or at least will be available to third parties under certain conditions (e.g. CC licenses, upon request, etc…); only a few of the considered datasets will not be publicly available due to copyright restrictions or the need to respect data protection laws.

The MOVING Data Management Plan is a working document that evolves during the lifetime of the project. For this reason, although no formal update of the deliverable is foreseen, one or more updated versions of the Data Management Plan will be produced as the project progresses and in accordance with the project's needs; these will be made available as updates to the original deliverable via the project's website.

Abbreviations

Abbreviation	Explanation
BTC	Billion Triple Challenge
CC	Creative Commons license
CHIME	Center for Information Mining and Extraction
DCC	Digital Content Creation
DCNNs	Deep Convolutional Neural Networks
DERI	Digital Enterprise Research Institute
DMP	Data Management Plan
DyLDO	Dynamic Linked Data Observatory
IPR	Intellectual property rights
JSON	JavaScript Object Notation
LDC	Linguistic Data Consortium
LOD	Linked Open Data
NIST	National Institute of Standards and Technology
OCLC	Online Computer Library Center
RCV1	Reuters Corpus, Volume 1
RDF	Resource Description Framework
TRECVID MED	TREC Video Retrieval Evaluation Multimedia Event Detection
TRECVID SIN	TREC Video Retrieval Evaluation Semantic Indexing

Abbreviation	Explanation
TSV	Tab-Separated Values
W3C	World Wide Web Consortium

1 Applied Methodology

The applied methodology for drafting the Data Management Plan of the project was based on the guidelines of the EC[1] and the DMP online tool[2], which can be used for implementing such a plan in a structured manner via a series of questions that need to be clarified for each dataset of the project. According to these guidelines, the Data Management Plan of MOVING addresses the points below on a dataset-by-dataset basis, reflecting the current status within the consortium about the data that will be produced:

- Dataset reference and name.
- Dataset description.
- Standards and metadata.
- Data sharing.
- Archiving and preservation (including storage and backup).

A more detailed description of the information that is considered and reported for each one of these subjects is provided in the following subsections.

1.1 Dataset reference and name

For convenient referencing of the data that will be collected and/or generated in the project we had to define a naming pattern. A referencing approach that contains information about the WP that owns/uses the dataset, the serial number of the dataset and the title of the dataset is the following: *MOVING_Data_"WPNo."_"DatasetNo."_"DatasetTitle"*. According to this pattern, an example dataset reference name could be *MOVING_Data_WP1_1_UserGeneratedContent*.

1.2 Dataset description

The description of the dataset that will be collected and/or generated includes information regarding the origin (in case of data collection), nature and scale of the data, as well as details related to the potential users of the data. Moreover, the description clarifies whether these datasets are expected to support a scientific publication, while information on the existence (or not) of similar data and the possibilities for integration and reuse is provided.

1.3 Standards and metadata

This section outlines how the data will be collected and/or generated, and which community data standards (if any) will be used at this stage. Moreover, it provides information on how the data will

[1] https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[2] https://dmponline.dcc.ac.uk

be organized during the project, mentioning for example naming conventions, version control and folder structures. For a detailed overview of the used standards the following questions were considered:

- How will the data be created?
- What standards or methodologies will be used?
- Which structuring and naming approach will be applied for folders and files?
- How different versions of a dataset will be easily identifiable?

In addition this section reports the types of metadata that will be created to describe the data and aid their discovery. How this information will be created/captured and where it will be stored is also reported. The aspects below were examined for determining the necessary ways and types of generating and using metadata:

- How these metadata are going to be captured/created?
- Can any of this information be created automatically?
- What metadata standards will be used and why?

1.4 Data sharing

This section describes how the collected and/or generated data will be shared. For this, it reports on access procedures and embargo periods (if any), and lists technical mechanisms and software/tools for dissemination and exploitation/re-use of these data. Moreover it determines whether access will be widely open or restricted to specific groups (e.g. due to participant confidentiality, consent agreements or Intellectual Property Rights (IPR), while it outlines any expected difficulties in data sharing, along with causes and possible measures to overcome these difficulties. In case a dataset cannot be shared, the reasons for this are mentioned (e.g. ethical rules of personal data and privacy-related considerations, intellectual property and commercial interests). Last but not least, identification of the repository where data will be stored, indicating in particular the type of repository (institutional, standard repository for the discipline, etc.) is also performed. The questions below were studied for concluding to the most appropriate sharing policy for each dataset of the project:

- How these data are going to be available to others?
- With whom will be the data shared, and under what conditions?
- Are any restrictions on data sharing required (e.g. limits on who can use the data, when and for what purpose)?
- What restrictions are needed and why?
- What actions will be taken to overcome or minimize restrictions?
- Where (i.e. in which repository) will the data be deposited?

1.5 Archiving and preservation

The established data archiving and preservation policy defines the procedures that will be put in place for long-term preservation of the data. In particular it indicates how long the data will be preserved and what is their approximate end volume. It also outlines the plans for preparing and documenting data for sharing and archiving. In case of not using an established repository, the Data Management Plan describes the resources and systems that will be put in place to enable the data to be curated and used effectively beyond the lifetime of the project.

A set of questions that were considered for defining the archiving and preservation policy for the datasets of the project is given below:

- What is the long-term preservation plan for the dataset (e.g. deposit in a data repository)?
- Are there sufficient resources, including storage and other equipment, to carry out this plan, or are any additional resources needed?

2 Datasets in MOVING

This section lists the datasets that will be created or collected for the needs of the MOVING project, grouping them per WP (with the exception of WP4, which is focused on the software development of the platform, thus is not expected to generate any new dataset, or use datasets other than those already specified by the other work packages). Based on the methodology presented in Section 1, each dataset is defined by: (a) its name, (b) its description, (c) the used standards and accompanying metadata, (d) the applied data sharing policy, and (e) the adopted mechanisms for its archiving and preservation. As a key component for the creation and management of these datasets, data privacy issues will be closely monitored from the beginning of the project by the entire consortium and the project's Data Protection Officer (Mr. Sebastian Gottfried from TUD), to ensure that the collection, use and sharing of the data will not raise ethical concerns. Finally, we should stress that all our plans for the collection, retention, storage and sharing of datasets in the project, as described in the present document, do comply with the applicable national and EU legislations.

2.1 WP1 Datasets

Dataset name	**MOVING_Data_WP1_1_Science2.0SurveyResults**
Dataset description	The dataset involves empirical data from the Science 2.0-Survey on the usage of online-based tools and Social Media in academia, conducted in 2014. The survey evaluated sociodemographic characteristics, the usage of 18 different online-based tools and Social Media for research, teaching, administration and communication, reasons for usage and non-usage and attitudes against the usage of online-based tools and Social Media (privacy concerns, computer anxiety and self-efficacy, curiosity). At the time of the evaluation of the data no similar data existed.
Standards and metadata	The survey was conducted as Online-Survey with the Software Questback/Unipark. • Number of units: 2083. • Number of Variables: 476. • Analysis Systems: SPSS, Stata. The attitudinal evaluation in this study is based on the following scales for measuring the acceptance of technology: • Privacy concerns (based on Xu, H.; Dinev, T.; Smith, J. & Hart, P. (2011): Information Privacy Concerns: Linking Individual Perceptions with Institutional Privacy Assurances. Journal of Association for Information Systems 12/12, S. 798–824).

	• Computer anxiety and self-efficacy (based on Venkatesh, V. & Bala, H. (2008): Technology Acceptance Model 3 and a Research Agenda on Interventions. Decision Sciences 39/2, S. 273–315). • Curiosity (based on Kashdan T.B., Rose P., Fincham F.D. (2004): Curiosity and exploration: Facilitating positive subjective experiences and personal growth opportunities. Journal of Personality Assessment 82, S. 291–305).
Data sharing	Dataset, codebook and method report are already published open access on GESIS Data Catalogue DBK, available via *doi:10.4232/1.12262*.
Archiving and preservation	The dataset is stored persistently on file servers of the GESIS data archive according to latest technical standards for long-term preservation. File servers are protected by applying the commonly used security measures for preventing unauthorized access and ensuring that security software is up-to-date with the latest released security patches. Preservation will be ensured by regularly backups of the original databases or file systems.

Dataset name	**MOVING_Data_WP1_2_EScienceResearchNetworkInterviewResults**
Dataset description	Within the ESF-funded project eScience research network, between December 2012 and June 2013, 19 semi-structured interviews have been held to evaluate changes of scientific practices in research, science administration, science communication and teaching. At the time of the evaluation in 2012 no similar dataset existed. The dataset will be useful for the requirement analysis of the platform (Task 1.1). There will be no possibility for reuse of the original data for other researchers (see below the data sharing conditions for this dataset).
Standards and metadata	The dataset consists of 19 anonymized interview transcripts (doc- or rtf-files). The interviews have been conducted as semi-structured interviews with researchers from universities in Saxony, including disciplines from the sciences, social sciences as well as engineering sciences and the status groups PhD, professor and research assistant.
Data sharing	Access to the interview dataset will be restricted to the MOVING project partners to ensure privacy protection and no abuse of the data. Results of the secondary analysis within WP1 will be made accessible to third parties via scientific publications, posters, project reports etc.

Archiving and preservation	The interview dataset is permanently stored on file servers of TUD. File servers are protected by applying the commonly used security measures for preventing unauthorized access and ensuring that security software is up-to-date with the latest released security patches. Preservation will be ensured by regularly backups of the file systems. Backups will be either conducted internally with the local administrator of the Media Center or with the help of a service partner within TUD, that offers free of charge backups with IBM Tivoli Storage System (TSM) including a guaranteed maximum 10 year time span to preserve this data. Additionally duration may be discussed with said partner.

Dataset name	**MOVING_Data_WP1_3 _Interviews**
Dataset description	The dataset consists of 10 anonymized interview transcripts with young researchers (PhD and Master students). The interviews consist of questions about: • The behavior towards research information searching and information management strategies. • The usage of online-based tools for research. • Training behavior and usage of online-based training services. The interviews are conducted because at the moment no similar data exists. The dataset will be useful for the requirement analysis of the platform (Task 1.1). This data may only be reused for research purposes under the data sharing conditions mentioned below.
Standards and metadata	The dataset consists of anonymized interview transcripts (doc or rtf files). The interviews are conducted as structured interviews.
Data sharing	Access to the interview dataset will be restricted to the MOVING project partners to ensure privacy protection and no abuse of the data. Results of the empirical analysis will be made accessible to third parties via scientific publications, posters, project reports etc.
Archiving and preservation	The interviews dataset is permanently stored on file servers of TUD. File servers are protected by applying the commonly used security measures for preventing unauthorized access and ensuring that security software is up-to-date with the latest released security patches. Preservation will be ensured by regularly backups of the file systems. Backups will be either conducted internally with the

	local administrator of the Media Center or with the help of a service partner within TUD, that offers free of charge backups with IBM Tivoli Storage System (TSM) including a guaranteed maximum 10 year time span to preserve this data. Additionally duration may be discussed with said partner.

Dataset name	**MOVING_Data_WP1_4_LabUserStudies**

Dataset description	Contains data collected from the user studies in the lab. The goal of such studies is to verify whether the requirements of the platform are met. In order to do so we will collect:
	• Eye-tracking data includes:
	  ○ 'Areas of Interest' (AOI) looked at. Typically such a record contains a user identifier, the timestamp and a Boolean value indicating whether a specific AOI has been looked at.
	  ○ Scanpath and fixation data includes a user identifier, a timestamp, the coordinates on the fixation and the duration of the fixation.
	  ○ Graphical visualization of data. Typical visualizations are heatmaps and scanpaths.
	• Log data from user interaction: user, URL, timestamp, event, object of the event.
	• Questionnaire responses.
	• Transcriptions of the think-aloud method and semi-structured interviews.
	• Outcomes of the study in terms of efficiency (completion time), effectiveness and satisfaction.
	This dataset is key to accomplish Task 1.4 and verify whether users' expectations with the MOVING platform are met. Some other parties may find it useful for reproducibility and repeatability reasons.

Standards and metadata	Most of the above follow a CSV format, except for the transcriptions that will be in TXT or RTF.
Data sharing	Open to the consortium since its creation, and open to all once the corresponding papers have been submitted for publication; except for log data, which gets stored in TUD servers and cannot leave them due to the German Data Protection Law.

Archiving and preservation	Data will be uploaded into the project's Zenodo[3] profile. Also, it will be stored on the data servers of the "Interaction Analysis and Modeling Lab" at http://iam-data.cs.manchester.ac.uk/data_files. This server automatically backups its contents to another machine through a CRON job.

2.2 WP2 Datasets

Dataset name	**MOVING_Data_WP2_1_EYUserStudyResultKC**
Dataset description	This dataset will contain data collected during the on-site user study of the KC at EY. It will consist of a set of filled-in questionnaires, interviews (with audio recordings), work observation data and workshops in order to extract detailed information about how auditors perform their work. The goal of the data is to discover how adaptive training and reflection support within the MOVING platform can be meaningfully implemented. The number of participants is not fixed yet. The collected data is useful and reusable for all MOVING partners who might be interested in how EY works and for the development of tools related to the MOVING platform tailored to the EY auditors' needs.
Standards and metadata	The questionnaire results (either collected with an online survey or paper-based) will be stored in EXCEL sheets and analyzed with SPSS or R, and then stored in the corresponding formats. The audio recordings of the interviews will be stored as typical .WAV files, their transcriptions will be stored in text files (.doc, .docx or txt files). Observation data will be noted down in text files, and photos (if allowed to be taken) will be stored in a typical image format like e.g. JPG, PNG, or TIF. Workshop information will also be stored in text files. All data will be collected at EY during different visits. None of this data will be collected automatically.
Data sharing	The access to the anonymized dataset will be restricted to the MOVING partners. In addition, no personalized data will be made available at any point in time due to the restrictions of the Austrian law (§ 46 DSG 2000 - Datenschutzgesetz). The data is used solely for research purposes and may not be copied and re-used for any other purpose. Fully anonymized aggregated analysis results may be included in public

[3] https://zenodo.org/collection/user-moving-h2020

	reports, deliverables, and scientific publications.
Archiving and preservation	The paper-based data will be stored in a locked board at the Know-Center. All digital data will be permanently stored in an encrypted file on an internal server at the Know-Center.

2.3 WP3 Datasets

Dataset name	**MOVING_Data_WP3_1_TRECVID**
Dataset description	This dataset is provided by NIST to the participants of the TRECVID SIN and MED tasks. It will be used for developing technologies for video annotation with visual concept labels. The dataset is divided in two main parts. The first part consists of approx. 18,500 videos (354 GB, 1,400 hours) under a Creative Commons (CC) license, in MPEG-4/H.264 format, and it is typically partitioned into training (approx. 11,200 videos, 10 seconds to 6,4 minutes long; 210 GB, 800 hours total) and testing set (approx. 7300 videos, 10 seconds to 4,1 minutes long; 144 GB, 600 hours total) for video concept detection methods. The total number of concepts is 346, and the annotation of each of these videos is based on a pair of XML and TXT files; the XML file contains information about the shot segments of the video and the TXT file includes the shot-level concept-based annotation of the video via a number of positive and negative concept labels. Finally, a TXT file with metadata describing sets of relations between these concepts in the form of "concept A implies concept B" and "concept A excludes concept B", is also available. The second part is a collection of approx. 63,000 videos (736 GB, 2,520 hours) in MPEG-4/H.264 format, created by the Linguistic Data Consortium and NIST. It is used for the development of video event detection techniques and is divided in three subsets: (a) a training set with 3,000 (50 GB, 80 hours) positive or near-miss videos, and 5,000 (51 GB, 200 hours) background (i.e., negative) videos, (b) a validation set of 23,000 videos (272 GB, 960 hours), and (c) an evaluation set of 32,000 videos (363 GB, 1280 hours). The number of considered events is 20, and the ground truth for this collection is stored in CSV files. These files provide the event-based annotations of the videos by defining the list of positive or near-miss videos for each visual event.
Standards and metadata	The videos of this static dataset are in MPEG-4/H.264 format, while their annotations and metadata are in TXT, XML and CSV files. The generated results after

	processing this dataset (extracted features, if any; automatic annotation results) will be stored in XML, JSON and MPEG-7 formats. They will be accompanied by a document (a word file) containing metadata with sufficient information to: (a) link it to the research publications/outputs, (b) identify the founder and research discipline, and (c) appropriate key words to help users to locate the data.
Data sharing	This is a dataset created and provided to us by NIST, under specific conditions that are linked with the TRECVID benchmarking activity. Sharing of the dataset is regulated by NIST, and we will comply with their requirements. We are not allowed to further share this dataset with 3rd parties. The results of our processing of the dataset (automatic annotation results in XML, JSON or MPEG-7 formats) will be uploaded into the project's Zenodo profile, under the express conditions that the data is used solely for the purposes of evaluating concept detection algorithms and may not be copied and re-used for any other purpose.
Archiving and preservation	As stated above, a set of processing outcomes of the dataset will be made available into the project's Zenodo profile. The original dataset and the analysis results will be stored on the file servers of CERTH (protected by applying the commonly used security measures for preventing unauthorized access and ensuring that security software is up-to-date with the latest released security patches) and backup provisions will be made.

Dataset name	**MOVING_Data_WP3_2_ImageNET**
Dataset description	This dataset contains images of the online ImageNet collection, which is organized and managed by the Stanford and Princeton Universities. It will be used for building and training Deep Convolutional Neural Networks (DCNNs) for video concept detection. In particular, ImageNet is an image dataset organized according to the WordNet hierarchy (currently only the nouns); for each node of the hierarchy, related images (often several hundreds or thousands of them) are provided. The current dataset is the one released in fall 2011 and is an updated version of the initial collection. It contains approx. 15 million images in high resolution JPEG format, which are clustered in categories that correspond to 22,000 distinct concepts of the WordNet structure. Images of each concept are quality-controlled and human-annotated.
Standards and metadata	This static dataset is composed by images that are mainly in high resolution JPEG format. The created metadata after analyzing these images can be: (a) local features

	extracted from these images, that are stored in BIN of TXT files, and (b) the output of the trained DCNNs (i.e., the classification decision), which is stored in TXT files. These data will be accompanied by a document (a word file) containing metadata with sufficient information to: (a) link it to the research publications/outputs, (b) identify the funder and discipline of the research, and (c) appropriate key words to help internal users to locate the data.
Data sharing	The ImageNet dataset is freely available for non-commercial research and/or educational use, by following the procedure and adopting the terms of use that are described in the ImageNet website.
Archiving and preservation	The original dataset and the results of processing it will be stored on the file servers of CERTH (protected by applying the commonly used security measures for preventing unauthorized access and ensuring that security software is up-to-date with the latest released security patches) and backup provisions will be made. The archiving and preservation of this dataset are performed by the Stanford and Princeton Universities; MOVING will have no involvement in this process.

Dataset name	**MOVING_Data_WP3_3_SocialMediaWebRetrieval**
Dataset description	This dataset comprises of the web content that will be collected by the MOVING platform as part of Task 3.2. Content collection will be seeded by a user-defined set of topics. The dataset will include both text and multimedia content and will be collected from different social networks such as Twitter, Google+, Youtube etc., and from selected webpages, such as research project websites.
Standards and metadata	The collected dataset from social networks will be stored in JSON format, in a MongoDB and will be closely following the format of the original sources, while the dataset from webpages will be temporarily stored for indexing in html format.
Data sharing	Due to the limitations imposed by Twitter (republishing a tweet is not allowed, only linking[4] or embedding[5] it is permitted) and the other social media platforms to be accessed, access to the dataset will be restricted to MOVING partners. Fully

[4] https://support.twitter.com/articles/80586#

[5] https://support.twitter.com/articles/20169559#

	anonymized aggregated analysis results may be included in public reports, deliverables, and scientific publications.
Archiving and preservation	The dataset will be stored persistently (i.e. guaranteed until project's end and planned to be kept also after the project for an undefined period of time) on a TUD server (protected by applying the commonly used security measures for preventing unauthorized access and ensuring that security software is up-to-date with the latest released security patches). Preservation will be ensured by backup of the original databases or file systems.

Dataset name	**MOVING_Data_WP3_4_PublicationMetaData**
Dataset description	This dataset comprises about 2,8 million metadata records of social science publications in German speaking countries collected at GESIS. The dataset is available to the public for searching via the Sowiport portal (www.sowiport.de). The dataset will be updated periodically by GESIS. The dataset is useful for the TUD use case. To the best of our knowledge, there is no similar collection for German speaking countries.
Standards and metadata	Each record contains common metadata on scholarly publications such as author(s), title, abstract, keywords, classification, and bibliographical data. All records are indexed using controlled social science vocabulary. Cooperation partners upload data in custom format via ftp. Files are automatically parsed w.r.t. the partner-specific formatting, fed into the database and made available on the platform. In case of errors, manual correction is done at GESIS where it is possible. There is no versioning. Entries can be overwritten but no old versions are stored, nor is the number of changes recorded.
Data sharing	Due to agreements with cooperation partners the dataset can be made available to third parties under the express conditions that the data is used solely for research purposes and may not be copied and re-used for any other purpose. Bulk downloads can be initiated for the purposes of the project using the OAI API at *sowiport.gesis.org/OAI/Home.*

Archiving and preservation	The dataset is persistently stored at a GESIS Vufind server in XML format. The server is protected by applying the commonly used security measures for preventing unauthorized access and ensuring that security software is up-to-date with the latest released security patches. Preservation will be ensured by regularly backups of the original databases or file systems. Description of the sources and statistics can be viewed on the website: *sowiport.gesis.org/Database.* The full database schema can be handed out on upon request.

Dataset name	**MOVING_Data_WP3_5_OAFulltexts**
Dataset description	This dataset comprises 5,400 open access full texts of social science publications. The full texts are available via the SSOAR repository (www.ssoar.info). The full-text server SSOAR, which is maintained at GESIS – Leibniz Institute for the Social Sciences, collects and archives literature of relevance to the social sciences and makes it available in open access on the Internet in accordance with the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities. SSOAR primarily pursues the so-called “Green Road to Open Access” (OA) and sees itself as a secondary publisher of quality-controlled literature. SSOAR currently comprise about 36,000 full texts. Most of them underlie a deposit licence. Of these 5,400 full texts are openly available. The dataset is permanently updated by contributions of the social scientists. The dataset is useful for the TUD use case. To the best of our knowledge, there is no similar collection for German speaking countries.
Standards and metadata	All full texts are described by a set of common metadata, such as author(s), title, abstract, keywords, classification, and bibliographical data, and are indexed using controlled social science vocabulary. Scientists upload pdf and metadata or only metadata manually. Publication is checked manually at GESIS for appropriateness. In case of errors, manual correction is done at GESIS. Data is made available at a point in time specified by the author. There are no automatic means of uploading data. All entries have a version number. All versions are available on the portal.

Data sharing	Due to license issues, the dataset can be made available to third parties under the express conditions that the data is used solely for research purposes and may not be copied and re-used for any other purpose. Bulk downloads can be initiated by anyone using the OAI API at *www.ssoar.info/OAIHandler*
Archiving and preservation	The dataset is persistently stored at a GESIS DSpace server. The server is protected by applying the commonly used security measures for preventing unauthorized access and ensuring that security software is up-to-date with the latest released security patches. Preservation will be ensured by regularly backups of the original databases or file systems. Some general information is available on the website: www.ssoar.info/home/veroeffentlichen-auf-ssoar.html. The schema is not described on publicly available sites, but it can be derived easily when looking at the fields offered on the portal or in the OAI xml files.

Dataset name	**MOVING_Data_WP3_6_ProjectMetaData**
Dataset description	This dataset comprises about 56,000 metadata record on social science research projects. The dataset is available to the public via the SOFISwiki portal (https://sofis.gesis.org/sofiswiki/Hauptseite). The dataset is permanently updated by contributions of the social scientists. The dataset is useful for the TUD use case. To the best of our knowledge, there is no similar collection for German speaking countries. As regards DFG projects there is an overlap with the DFG project database.
Standards and metadata	Each record contains common metadata such as researcher(s), title, abstract, keywords, classification, and a number of metadata on research methods and publications. All records are indexed using controlled social science vocabulary. Scientists upload or modify metadata of their project entries manually. All edits will be checked manually at GESIS for appropriateness. In case of errors, manual correction is done at GESIS. The data are then made available to the public via the SOFISwiki portal. Additionally, bulk imports from cooperation partners are done regularly. Projects that have been provided by DFG are currently not being

	processed due to overlaps with the existing database. There is no decided versioning, but the display format of Semantic Mediawiki allows viewing of previous versions for single projects.
Data sharing	Due to the terms of use of the SOFISwiki (Creative Commons License CC BY-NC-SA 3.0 DE), the dataset can be made available to cooperation partners of GESIS within the framework of common projects under the express conditions that the data is used solely for research purposes and may not be copied and re-used for any other purpose. There is no repository for sharing the whole dataset, but it can be provided by GESIS upon request for the purposes of the MOVING project.
Archiving and preservation	The dataset is persistently stored at a GESIS Semantic MediaWiki server. The server is protected by applying the commonly used security measures for preventing unauthorized access and ensuring that security software is up-to-date with the latest released security patches. Preservation will be ensured by regularly backups of the original databases or file systems. There is no particular documentation of the project database. However the structure is fairly simple and can be derived when retrieving results on SOFISwiki.

Dataset name	**MOVING_Data_WP3_7_InteractionData**
Dataset description	This dataset will contain the data generated through the use of the MOVING platform. Consequently it will consist of interaction events at a different level of granularity: • Low-level events are those triggered through the use of the keyboard, mouse, touchscreen or mouse wheel. • Higher-level events are semantically more meaningful and will describe specific actions on the platform: “search for X”, “open Y” and similar. A row of this dataset will look like: *user, URL, timestamp, event, object of the event*. This dataset is key to accomplish Task 3.3 in order to infer knowledge acquisition from user behavior on the MOVING platform.
Standards and metadata	This dataset will be formatted as CSV files.

Data sharing	While this data is open for the members of the MOVING consortium, it cannot be shared, as it will be generated in TUD servers and due to German Data Protection Law, data should remain there. Data will be analyzed by remotely accessing TUD servers and will never be downloaded elsewhere.
Archiving and preservation	Due to the above reasons the data will be kept in TUD servers. The file servers themselves are protected by applying the commonly used security measures for preventing unauthorized access and ensuring that security software is up-to-date with the latest released security patches. Preservation will be ensured by regularly backups of the file systems. These backups will be conducted with the help of a service partner within TUD, that offers free of charge backups with IBM Tivoli Storage System (TSM) including a guaranteed maximum 10 year time span to preserve this data. Additionally duration may be discussed with said partner.

Dataset name	**MOVING_Data_WP3_8_BTC2014**
Dataset description	The Billion Triples Challenge 2014 (BTC)[6] dataset contains structured data from various domains like Government, Publications, Life sciences, User-generated content, Cross-domain, Media, Geographic and Social Web. The latest dataset from 2014 covers 47,560 different pay-level domains and contains 4,090,758,596 (4 Billion) RDF-quads in total. All quads are available in one dump with a volume of 1.1TB (unzipped) or 52GB (compressed). Besides the raw data, the BTC dataset comes with additional metadata information regarding origin and statistical information of the sources. The dataset is crawled from the public web via the open source program LDSpider[7]. Since the data is crawled from the Web, it is of varying quality regarding structure and content. The BTC2014 dataset is relevant for researchers analysing Linked Open Data (LOD) and developing LOD-based applications.
Standards and metadata	All data including the metadata is formatted in standards defined by the W3C, namely the RDF format and N-Quads.

[6] http://km.aifb.kit.edu/projects/btc-2014/

[7] https://github.com/ldspider/ldspider

Data sharing	The original BTC 2014 dataset is hosted by KIT in Karlsruhe and can be freely downloaded. The procedure is described on the website of the dataset.
Archiving and preservation	The dataset will be stored persistently at ZBW on dedicated servers. The servers are protected by established security measures for preventing unauthorized access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups.

Dataset name	**MOVING_Data_WP3_9_LODLaundromat**
Dataset description	The LOD Laundromat[8] provides access to a very large collection of Linked Open Data (LOD). The creators do not claim to provide a "new dataset, but rather a uniform point of entry to a collection of cleaned siblings of existing datasets." (Beek et al., 2014). The latest dump from 2016 contains 38,606,408,854 (38 Billion) RDF-quads in total. All quads are available in one dump with a volume of 291GB (gzipped N-Triples). The dataset is crawled from the public web via the open source program WashingMachine[9]. Since the data is crawled from the Web, it is of varying quality regarding structure and content. Although the data is cleansed, it does not mean that any kind of professional content review has been performed. The LODLaundromat dataset is relevant for researchers analysing Linked Open Data (LOD) and developing LOD-based applications.
Standards and metadata	All data including the metadata is formatted in standards defined by the W3C, namely the RDF format and N-Quads.
Data sharing	The original LOD Laundromat dataset can be accessed via the Wardrobe[10]. The procedure is described on the website of the dataset.
Archiving and preservation	The dataset will be stored persistently at ZBW on dedicated servers. The servers are protected by established security measures for preventing unauthorized access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups.

[8] http://lodlaundromat.org/

[9] https://github.com/LOD-Laundromat/LOD-Laundromat

[10] http://lodlaundromat.org/wardrobe/

Dataset name	**MOVING_Data_WP3_10_DyLDO**
Dataset description	The Dynamic Linked Data Observatory (DyLDO)[11] dataset contains weekly snapshots of larger crawls from the Linked Open Data (LOD) cloud starting with 06.05.2012. On average (over the first 29 datasets) there are 1,738.6 pay-level domains in the dataset with a total of 94,725,595 quads bibliography (Käfer et al., 2013). The weekly crawls are stored in separate dumps. As in the BTC2014 dataset, the DyLDO dataset also contains structured data from various domains. Although the snapshots are smaller compared to the BTC2014, they are of added value since they capture the evolution of the LOD cloud over time. Furthermore, also the Dynamic Linked Data Observatory uses the LDSpider to conduct the weekly crawls. The DyLDO dataset is relevant for researchers analysing the temporal evolution of Linked Open Data (LOD) and developing LOD-based applications considering the time aspect.
Standards and metadata	All data is formatted in standards defined by the W3C, namely the RDF format and N-Quads.
Data sharing	The dataset is provided by DERI in Ireland. The weekly crawls can be freely downloaded. The procedure is described on the website of the dataset.
Archiving and preservation	The dataset will be stored persistently at ZBW on dedicated servers. The servers are protected by established security measures for preventing unauthorized access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups.

Dataset name	**MOVING_Data_WP3_11_ZBWEconomicsDataset**
Dataset description	The ZBWEconomicsDataset is provided by the partner institution ZBW – Leibniz Information Centre for Economics. ZBW is running a search portal, called EconBiz[12], for economics' scientific publications. From EconBiz, ZBW obtained 1 million URLs of open access scientific publications and generated a dataset of about 413,000 full-texts in PDF format and the metadata (e.g., authors, title, year of publication). From

[11] http://swse.deri.org/dyldo/data/

[12] http://www.econbiz.de/

	the 413,000 scientific publications in the economics domain, 280,000 publications are in English. The remaining publications cover 41 other languages, most notably German, French and Spanish. In addition to the PDF format, the publications are also converted to plain ASCII text format (TXT). Furthermore, from the English publications we extracted about 200,000 images which are most likely scholarly figures. From these figures, we randomly selected and manually annotated 121 scholarly figures. The figures vary in type (e.g., bar chart, pie chart, map), quality and topic. Ground truth information containing the position, orientation for each text line is stored in TSV and JSON format. Information about the origin of the individual figures is provided in textual form in form of an identifier to the EconBiz portal. The scientific publications of the ZBWEconomicsDataset are relevant for researchers working on information retrieval, recommendation, and document classifications tasks. The annotated set of figures is relevant for researchers who try to improve the indexing of figure-like images based on the text inside these figures.
Standards and metadata	The full-texts are available in plain ASCII format (TXT) and PDF format. The metadata are provided in JSON format. The ground truth information for the manually annotated scholarly figures is available in TSV and JSON format.
Data sharing	The sharing of the data is controlled by ZBW. The dataset is available only to the participating partners of MOVING due to possible copyright restrictions that some collected publications might have. Furthermore, the dataset can be requested by other research institutions via individual agreements with ZBW.
Archiving and preservation	The dataset will be stored persistently on dedicated servers of ZBW. The servers are protected by established security measures for preventing unauthorized access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups.

Dataset name	**MOVING_Data_WP3_12_NYT**
Dataset	The New York Times dataset is organized and managed by the Linguistic Data Consortium (LDC)[13]. It contains a set of more than 1.8 million articles that have been

[13] https://www.ldc.upenn.edu

description	published between 1st January 1987 and 19th of June 2007. More than 650,000 articles contain a professional summary and more than 1.5 million articles have been manually annotated by librarians with various tags for places, people, organizations and topics. The dataset is available in form of XML files. Each file contains information about a single article. For instance[14] the publication date, section, author, title and contents. The dataset is relevant for researchers working on information retrieval, recommendation, and document classifications tasks.
Standards and metadata	A copy of the NYT dataset is available in form of XML documents in the NITF format. The data is provided together with an open source java tools for parsing the documents into memory objects.
Data sharing	The dataset is provided and managed by the Linguistic Data Consortium (LDC), The Trustees of the University of Pennsylvania. The dataset is individually licensed (per organization or individual) for some fee[15]. The data is owned by the New York Times Company.
Archiving and preservation	The dataset and analysis results will be stored persistently at ZBW on protected servers. The servers are protected by established security measures for preventing unauthorized access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups. The archiving and preservation of this dataset are performed be the LDC. MOVING will have no involvement in this process.

Dataset name	**MOVING_Data_WP3_13_Yago**
Dataset description	YAGO[16] is a knowledge base created by the Max Planck Institute in Saarbrucken and derived from Wikipedia[17], WordNet[18] and GeoNames[19]. The YAGO-NAGA project started in 2006 with the aim to build a "*highly accurate knowledge base of common facts in a machine-processible representation*". In 2012, the second version of YAGO

[14] https://catalog.ldc.upenn.edu/desc/addenda/LDC2008T19_large.jpg
[15] https://catalog.ldc.upenn.edu/LDC2008T19
[16] www.mpi-inf.mpg.de/YAGO/
[17] https://www.wikipedia.org/
[18] https://wordnet.princeton.edu/
[19] http://www.geonames.org/

	(YAGO2) has been released (Hoffart, J, Suchanek, F, M, Berberich, C, Weikum, G, 2013), followed by the third version (YAGO3) in 2015. YAGO3 contains more than 10 million entities with more than 120 million multilingual properties. YAGO's accuracy has been manually evaluated using a sample of facts and the extrapolated accuracy is between 90.84% and 99.22%[20]. The dataset is relevant for researchers working on entity extraction, entity classification, document retrieval and others.
Standards and metadata	YAGO is available in RDF/Turtle and TSV format.
Data sharing	YAGO is provided by Max Planck Institute for Informatics in Saarbrucken and is licensed under the CC 3.0 license.
Archiving and preservation	The dataset and analysis results will be stored persistently at ZBW on protected servers. The servers are protected by established security measures for preventing unauthorized access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups. The archiving and preservation of this dataset is done by the Max Planck Institute for Informatics in Saarbrucken together with the Databases and Information Systems Group[21] and the DBWeb team of Télécom ParisTech[22]. MOVING will have no involvement in this process.

Dataset name	**MOVING_Data_WP3_14_ScientificPublicationRetrieval**
Dataset description	The dataset will consist of a set of publicly available scientific publications that are obtained via established crawlers and protocols from corresponding open access archives. All publications will be associated with their domain (e.g. economics, computer science). The publications are stored in PDF format as well as in plain ASCII format (TXT). The OCLC OAI Harvester2[23] toolkit will be used for harvesting the

[20] http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/statistics/

[21] https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/

[22] http://dbweb.enst.fr/research/

[23] https://www.openarchives.org/OAI/2.0/guidelines-harvester.htm

	documents and keeping the crawl up-to-date. The crawled documents from the economics-related repositories (e.g. AgEcon[24] and Munich Personal RePEc[25]) Archive will support the EY use case, while the crawled data from the computer science related repositories (e.g. CiteseerXData and arXiv Bulk Data[26]) will support the TUD use case. The scientific publications of this dataset are relevant for researchers working on information retrieval, recommendation, and document classifications tasks.
Standards and metadata	The publications will be stored in PDF and plain ASCII (TXT). The metadata of the dataset will be formatted and stored in JSON and CSV formats. The metadata will include the following information such as the domain, title, authors, venue and year.
Data sharing	The publications are collected from open access repositories such as those mentioned above using standard protocols and tools like the OCLC's OAI Harvester2. The dataset is available only to the participating partners of MOVING due to possible copyright restrictions that some collected publications might have.
Archiving and preservation	The dataset and the analysis results will be stored persistently at ZBW on protected servers. The servers are protected by established security measures for preventing unauthorized access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups. The archival of the publications is performed by the crawled archive providers. MOVING will have no involvement in this process.

Dataset name	**MOVING_Data_WP3_15_DBPedia**
Dataset description	DBpedia is a Knowledge Base derived from Wikipedia. In 2007 the first version of DBpedia has been published. The project was initiated by the Free University of Berlin, the University of Leipzig[27] and OpenLink Software[28]. In the English version of the dataset, DBpedia currently has 4.22 million entities, including at least 1,445,000

[24] http://ageconsearch.umn.edu/
[25] https://mpra.ub.uni-muenchen.de/
[26] https://arxiv.org/help/bulk_data
[27] https://www.zv.uni-leipzig.de/en/
[28] http://www.openlinksw.com/

	people, 735,000 places, 123,000 music albums, 87,000 films and 19,000 video games that are described and licensed under a creative commons (CC) license. The dataset is stored using the Resource Description Framework (RDF) and queries can be made using SPARQL[29]. The dataset is relevant for researchers working on entity extraction, entity classification, document retrieval and others.
Standards and metadata	The DBPedia dataset is not only available in RDF, but also as a JSON and CSV tabular version. Since DBPedia extracts information from Wikipedia, The dataset provides structured information about the articles in forms of RDF Triples. For instance, the birth date of a person will take the form (person, date of birth, date).
Data sharing	The textual content is reusable in the terms of the GNU Free Documentation License (GFDL) and the Creative Commons Attribution-Share-Alike 3.0 License.
Archiving and preservation	The dataset and the analysis results will be stored persistently at ZBW on protected servers. The servers are protected by established security measures for preventing unauthorized access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups. The archiving and preservation of this dataset are performed by the Free University of Berlin, the University of Leipzig and OpenLink Software. MOVING will have no involvement in this process.

Dataset name	**MOVING_Data_WP3_16_RCV1**
Dataset description	The Reuters' dataset, known as "Reuters Corpus, Volume 1" or RCV1 is provided by NIST for research purposes. The RCV1 contains more than 800.000 manually-labelled Reuters' news stories. All news stories were written in English and published between 20th of August 1996 and 19th of August 1997. The news stories are organized in files. Each news article is organized in a separate XML file. Each XML file contains some information about an article. For instance title, publisher, contents and location. The news articles are categorized and controlled using the following three main vocabularies[30]: (a) Topics (126 topics), (b) Industries (870) and (c) Regions (366 geographic codes). All documents are contained in one dump with

[29] http://wiki.dbpedia.org/OnlineAccess

[30] http://jmlr.csail.mit.edu/papers/volume5/lewis04a/lewis04a.pdf

	a volume of 2.5GB. The dataset is relevant for researchers working on information retrieval, recommendation, and document classifications tasks.
Standards and metadata	The dataset files have been formatted in XML.
Data sharing	The copyright of the dataset is reserved to Reuters Ltd and/or Thompson Reuters. The dataset is provided and regulated by NIST. It can be used for research purposes. Every user or organization has to request his/its own copy[31][32].
Archiving and preservation	The dataset and analysis results will be stored persistently at ZBW on protected servers. The servers are protected by established security measures for preventing unauthorized access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups. The archiving and preservation of this dataset are performed by NIST. MOVING will have no involvement in this process.

Dataset name	**MOVING_Data_WP3_17_EnglishLanguageWikimedia**
Dataset description	Wikimedia Foundation, the parent organization of Wikipedia, published a dump[33] of more than 4.4 million Wikipedia articles containing more than 1.9 billion words. Wikimedia updates the dataset regularly, roughly twice a month. A free copy of this dataset is available in form of XML and SQL files. Each file contains information about an article. For instance the publication date, section, author, title and contents.[34] The dataset is relevant for researchers working on information retrieval, recommendation, and document classifications tasks.
Standards and metadata	The dataset is available in form of XML and/or SQL files.
Data sharing	All text contents can be used in the terms of the GNU Free Documentation License

[31] http://trec.nist.gov/data/reuters/ind_appl_reuters_v4.html

[32] http://trec.nist.gov/data/reuters/org_appl_reuters_v4.html

[33] https://dumps.wikimedia.org/

[34] https://catalog.ldc.upenn.edu/desc/addenda/LDC2008T19_large.jpg

	(GFDL) and the Creative Commons Attribution-Share-Alike 3.0 License.
Archiving and preservation	The dataset and analysis results will be stored persistently at ZBW on protected servers. The servers are protected by established security measures for preventing unauthorized access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups. The archiving and preservation of this dataset are performed by the Wikimedia Foundation. Wikimedia usually updates the dataset twice a month. MOVING will have no involvement in this process.

Dataset name	**MOVING_Data_WP3_18_CHIME**
Dataset description	The CHIME[35] dataset was created by the Center for Information Mining and Extraction, School of Computing, National University of Singapore (NUS). It consists of two sets of figures, one based on real world data created by Yang et al.[36] (Yang, L et al. 2006) and one based on synthetic data created by Jiuzhou[37], each having different numbers of bar charts, pie charts and line graphs. The synthetic set contains 85 figures while the other subset contains 115 figures. It was published to promote research activities involving figures. Each figure, from the dataset created by NUS, is accompanied with ground truth information in plain ASCII (TXT) and XML format which contain the position and content of all text and graphic elements inside the figure, but no information about their orientation. A transformed set of ground truth information about the text content in TSV and JSON format was created by ZBW and is available similar to the Economics Figures dataset. The annotated set of figures is relevant for researchers who try to improve the indexing of figure-like images based on the text inside these figures.
Standards and metadata	For representing the metadata and gold standard the TSV and JSON formats are used.

[35] https://www.comp.nus.edu.sg/~tancl/ChartImageDataset.htm

[36] https://www.comp.nus.edu.sg/~tancl/publications/c2006/das06-062.pdf

[37] https://www.comp.nus.edu.sg/~tancl/ChartImageDatabase/Report_Zhaojiuzhou.pdf

Data sharing	The dataset is publicly available.[38] The datasets with the extended ground truth information provided by ZBW are also publicly available.[39]
Archiving and preservation	The dataset and analysis results will be stored persistently at ZBW on protected servers. The servers are protected by established security measures for preventing unauthorized access and ensuring that security software is regularly updated and the latest security patches are applied. The dataset is contained in daily professional backups. The archiving and preservation of the figure dataset is performed by NUS. MOVING will have no involvement in this process. The archiving and preservation of the extended gold standard is conducted by ZBW and the same mechanisms are applied as described above for storing and managing the dataset and analysis results.

2.4 WP5 Datasets

Dataset name	**MOVING_Data_WP5_1_MOVINGPublications**
Dataset description	This dataset will contain manuscripts reporting the conducted scientific work in MOVING, which have been accepted for publication in high-quality peer-reviewed journals and conferences. All these publications will include a statement with acknowledgement to the MOVING project, while their content may vary from the description of specific analysis techniques, to established evaluation datasets and individual components or parts of the MOVING platform.
Standards and metadata	Most commonly, these documents will be stored in PDF format. Each document will be also accompanied by: (a) a short description with the abstract of the publications, (b) the LaTeX-related BIB file with its citation, and (c) details about the venue (e.g. conference, workshop or benchmarking activity) or journal where it was published. This dataset will be extended whenever new submitted works are accepted for publication in conferences or journals. A simple log file of the performed updates of the dataset will be maintained by CERTH in the project wiki (hosted by a CERTH server).
Data sharing	This dataset will be publicly available, following the guidelines of the EC[40] for open

[38] https://www.comp.nus.edu.sg/~tancl/ChartImageDataset.htm

[39] http://www.kd.informatik.uni-kiel.de/en/research/software/text-extraction

	access to scientific publications and research data in Horizon2020.
Archiving and preservation	Self-archiving (also known as "green open access") will be applied for ensuring open access to these publications. According to this archiving policy the author(s) of the publication will archive (deposit) the published article or the final peer-reviewed manuscript in online repositories: (a) personal webpage(s), (b) the project website[41], and (c) into the project's Zenodo profile, after its publication. Nevertheless, the employed archiving policy will be fully aligned with restrictions concerning embargo periods that may be defined by the publishers of these publications, making the latter publicly available in certain repositories only after their embargo period has elapsed.

Dataset name	**MOVING_Data_WP5_2_MOVINGPresentations**
Dataset description	This dataset will consist of presentations prepared for reporting MOVING-related scientific work or progress made, in a variety of different events, such as conferences, workshops, meetings, exhibitions, interviews and so on.
Standards and metadata	Most commonly these presentations will be in PPT or PDF format. Information related to: (a) the authors, (b) the presenter, (c) the venue and (d) the date of the presentation will be also stored in plain text. This dataset will be extended whenever new MOVING presentations are prepared and publicly released. A simple log file of the performed updates of the dataset will be maintained by CERTH in the project wiki (hosted by a CERTH server).
Data sharing	The project presentations will be made publicly available after their presentation at the venue/event they were prepared for.
Archiving and preservation	The project presentations will be publicly available for view and download via the SlideShare channel of the project[42] and the project's Zenodo profile, while links to

[40] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

[41] http://moving-project.eu/index.php/publications

[42] http://www.slideshare.net/MOVING_EU

	the presentations on SlideShare will be also added to the relevant webpage of the project website[43]. The latter is hosted in a CERTH server that is protected by applying the commonly used security measures for preventing unauthorized access and ensuring that security software is up-to-date with the latest released security patches.

Dataset name	**MOVING_Data_WP5_3_MOVINGSoftwareDemosAndTutorials**
Dataset description	This dataset will collect information regarding the developed and utilized MOVING technologies. Public video demonstrations, tutorials with instructions of use, documentations as well as links to publicly-released online instances of these technologies will be also included.
Standards and metadata	A variety of different formats will be used for storing the necessary information. In particular, video demonstrations can be (but not limited to) MP4, AVI or WEBM files, software tutorials and documentations can be written in PDF format, online documentations of tools and services can be presented in plain text, and presentations can be stored in PPT or PDF format. This dataset will be extended whenever new content related to the MOVING developed technologies (e.g. video/web demos, tutorials, documentation) is prepared and publicly released. A simple log file of the performed updates of the dataset will be maintained by CERTH in the project wiki (hosted by a CERTH server).
Data sharing	Information related to the developed MOVING technologies, including video demonstrations, documentations, presentations and tutorials with instructions of use, will be publicly available supporting the dissemination of the project's activities and the exploitation of the project's outcomes. However, confidentiality control will be applied on each piece of information in order to avoid the release of inappropriate information that could have a negative impact to the project's progress and developments.
Archiving and preservation	Data related to the developed MOVING technologies, tools and applications will be archived and made publicly available through the relevant webpage of the project website[44], which is hosted by a CERTH server that is protected by applying the

[43] http://moving-project.eu/index.php/presentations

[44] http://moving-project.eu/index.php/tools-and-services

	commonly used security measures for preventing unauthorized access and ensuring that security software is up-to-date with the latest released security patches. Moreover, the created video demos and tutorials will be also available for view via the YouTube channel of the MOVING project[45].

Dataset name	**MOVING_Data_WP5_4_MOVINGNewsletter**
Dataset description	This dataset will comprise the released newsletters for disseminating the activities and the progress made in the MOVING project.
Standards and metadata	The newsletters will be prepared and stored in PDF format, while information regarding their release date will be provided. This dataset will be extended whenever new project newsletters are publicly released. A simple log file of the performed updates of the dataset will be maintained by CERTH in the project wiki (hosted by a CERTH server).
Data sharing	The newsletters of the project will be publicly available online right after their official release.
Archiving and preservation	An online archive with open access to the released newsletters of the project will be maintained at the relevant webpage of the project website[46], which is hosted by a CERTH server that is protected by applying the commonly used security measures for preventing unauthorized access and ensuring that security software is up-to-date with the latest released security patches.

Dataset name	**MOVING_Data_WP5_5_MOVINGPlatformProspects**
Dataset description	This dataset comprises the opinions expressed by the MOVING project staff and stakeholders concerning the future prospects of technologies, business models and socio-economic trends relevant for the sustainability for the platform. The data will be gathered within a Delphi survey to be performed in Task 5.2.

[45] https://www.youtube.com/channel/UCLpMLXQQaHDv0CJMG5Sc7mg

[46] http://moving-project.eu/index.php/newsletters

Standards and metadata	This dataset will be semi-structured: the quantitative opinions will be stored in a MySQL database. Each record will contain the metadata: an identifier of the respondent, and date the opinion(s) was expressed. The opinions themselves will be the main body of the database consisting of quantitative expressions concerning the state of technology, social, business or economic environment of the MOVING platform at several prescribed moments in the future, the relations between different elements of these environments, and narrative descriptions and justifications. This dataset will also contain the results of verifications and trust analysis of expressions contained therein.
Data sharing	Access to the dataset will be initially restricted to MOVING partners; it will be made publicly available only after a peer-reviewed publication of the analysis of this data by MOVING project staff is published or accepted for publication. The above sharing policy is selected because raw data may be misused by inexperienced readers, specifically by using inappropriate statistical analysis methods, too high or too low confidence values, or inappropriate statistical tests. Therefore, false conclusions may be drawn from the survey data by authors who do not bear full responsibility for gathering and processing the dataset. The use of the data gathered from experts will require specialist, carefully and appropriately selected statistical methods, fully compatible with all the survey process. If published results of an inadequate statistical analysis of this dataset are not accompanied by full methodological details, which is a frequent malpractice in scientific community, a proof that the facts are different from those presented by authors from outside of the project may be very difficult or considerably delayed if there were no possibility to refer to the correct analysis published by MOVING authors before.
Archiving and preservation	The dataset will be archived at a PBF data server in a MySQL database (structured replies to the survey) and as text notes (free opinions expressed by the survey participants during collaborative sessions supplementing the survey) during the project and its durability period of 5 years after the project end. After the publication, the dataset may also be uploaded into the project's Zenodo profile (the data will be exported in .csv or .xls format), and will also be stored on MOVING project website as a supplementary material to project publications. A comprehensive description will be contained in the publications with data analysis results.

2.5 WP6 Datasets

Dataset name	**MOVING_Data_WP6_1_MOVINGDeliverables**
Dataset description	This dataset will be composed of the project deliverables that have to be prepared and submitted to the EC during the project's lifetime, according to the contractual obligations of the MOVING consortium.
Standards and metadata	These documents will be stored in PDF format. For each deliverable we will provide: (a) the list of authors, (b) a brief description of its content (i.e. its executive summary), (c) the related WP of the project, and (d) the contractual date for their submission to the EC. This dataset will be extended whenever new deliverables are submitted to the EC. A simple log file of the performed updates of the dataset will be maintained by CERTH in the project wiki (hosted by a CERTH server).
Data sharing	The public project deliverables will be made publicly available after their submission to the EC, via the project's website, while an abstract will be published for the confidential deliverables.
Archiving and preservation	This dataset will be maintained on the project wiki and the relevant webpage of the project website[47], both hosted by a CERTH server which is protected by applying the commonly used security measures for preventing unauthorized access and ensuring that security software is up-to-date with the latest released security patches. This webpage will grant open access to the PDF file of each listed public deliverable.

[47] http://moving-project.eu/index.php/deliverables

3 Conclusions

The Data Management Plan by the members of the consortium of the MOVING project was presented in this deliverable. This plan involves every dataset that will be collected, processed or generated during the lifetime of the project. In the present document we discussed 29 different datasets, both pre-existing ones and newly-created within MOVING. Most of these datasets are or will be made publicly available, to the benefit of the broader scientific community. The present document represents the dataset-related status and planning at month 6 of the MOVING project; as such, it may change to some extent during the remaining lifetime of the project, by e.g. the use or generation of new datasets that are not currently foreseen. To account for such changes, which are normal and expected within a 36-month project, we foresee that one or more updated versions of the Data Management Plan will be produced as the project progresses and in accordance with the project's needs. Though not being formal deliverables of the project, these updates to the Data Management Plan will be made available via the project's website.

References

Käfer, T, Abdelrahman, A, Umbrich, J, O'Byrne, P, Hogan, A (2013). “Observing Linked Data Dynamics” in Proc. of the 10th Extended Semantic Web Conference, ESWC, Montpellier, France, pp. 213-227.

Hoffart, J, Suchanek, F, M, Berberich, C, Weikum, G (2013). ”YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia”, Artificial Intelligence, Vol. 194, pp.28-61.

Yang, L, Huang, W, Tan, Ch, L, (2006). “Semi-automatic Ground Truth Generation for Chart Image Recognition“, in Proc of the 7th International Workshop (Document Analysis Systems VII), DAS, Nelson New Zealand, pp.324-335.